Parallel hierarchical sampling: a general-purpose class of multiple-chains MCMC algorithms

Fabio Rigat*

Antonietta Mira†

September 29th, 2009

Abstract

This paper introduces the Parallel Hierarchical Sampler (PHS), a class of Markov chain Monte Carlo algorithms using several interacting chains having the same target distribution but different mixing properties. Unlike any single-chain MCMC algorithm, upon reaching stationarity one of the PHS chains, which we call the "mother" chain, attains exact Monte Carlo sampling of the target distribution of interest. We empirically show that this translates in a dramatic improvement in the sampler's performance with respect to single-chain MCMC algorithms. Convergence of the PHS joint transition kernel is proved and its relationships with single-chain samplers, Parallel Tempering (PT) and variable augmentation algorithms are discussed. We then provide two illustrative examples comparing the accuracy of PHS with

*Department of Statistics and Centre for analytical Science, University of Warwick, UK; f.rigat@warwick.ac.uk

†Professor of Statistics, Department of Economics, University of Insubria, Italy; antonietta.Mira@uninsubria.it

that of various Metropolis-Hastings and PT for sampling multimodal mixtures of multivariate Gaussian densities and for 'banana-shaped' multivariate distributions with heavy tails. Finally, PHS is applied to approximate inferences for two Bayesian model uncertainty problems, namely selection of main effects for a linear Gaussian multiple regression model and inference for the structure of an exponential treed survival model.

Keywords: Bayesian model selection, classification and regression trees, Gaussian mixtures, heavy tails, linear regression, Metropolis-Hastings algorithm, multimodality, multiple-chains Markov chain Monte Carlo methods, survival analysis.

Introduction

Let $\theta \in \Theta$ be a random variable with cumulative distribution function $F(\theta)$ and probability density or probability mass function $f(\theta)$. Also let $K_f(\theta_i, \theta_{i+1})$ be a transition kernel defining the probability that a Markov chain, having state-space Θ and with $f(\theta)$ as its target distribution, jumps from a current state θ_i to a new state θ_{i+1}. Markov chain Monte Carlo (MCMC) algorithms generate sequences of dependent draws $\{\theta_i\}_{i=1}^N$ having $f(\theta)$ as their stationary distribution (Gelfand and Smith [1990], Smith and Roberts [1993], Neal [1993], Gilks et al. [1995b], Gamerman [1997], Robert and Casella [1999] and Liu [2001]). A sufficient set of conditions for convergence fulfilled by almost all current MCMC algorithms is that $K_f(\theta_i, \theta_{i+1})$ is reversible, F-irreducible and aperiodic, implying that $f(\theta)$ is its unique stationary and limiting distribution (Nummelin [1984], Robert and Casella [1999]) and that the strong law of large numbers and the central limit the-

orem hold for any function $g(\cdot) \in L^2(F)$ (Tierney [1994]).

These MCMC methods pioneered by Metropolis and Ulam [1949] and Metropolis et al. [1953] have been progressively adopted and adapted for a wide range of inferential problems, recently including challenging applications in phylogenetic inference (Yang and Rannala [1997], Mau et al. [1999], Li et al. [2000], Huelsembeck et al. [2002] Huelsembeck et al. [2004], Lunter et al. [2005]), molecular simulation (Lin et al. [2003]), DNA sequence alignment (Husmeier and McGuire [2003]) and discovery of gene regulatory networks (Li [2006]). Recent interest in MCMC methods for genetic model selection problems has also emphasized limitations of standard algorithms such as the Gibbs sampler and Metropolis-Hastings leading to poor mixing over large model spaces (Mossel and Vigoda [2005], Mossel and Vigoda [2006], Waagepetersen et al. [2008], Lakner et al. [2008]). Potential limitations to mixing of these algorithms when their stationary distributions are multi-modal or have fat tails are in fact well-known in the statistics literature (Cappé and Robert [2000], Celeux et al. [2000]), echoing long-standing research on their convergence properties (Rosenthal [1995], Robert [1995], Roberts and Tweedie [1996b], Athreya et al. [1996], Mengersen and Tweedie [1996], Cowles and Carlin [1996], Brooks [1998], Brooks and Roberts [1998], Roberts and Rosenthal [1998b], Jarner and Roberts [2002], Flegal et al. [2008]). These limitations have in part motivated the development of alternative MCMC sampling strategies using tempered distributions (Geyer and Thompson [1995], Neal [1996], Liang and Wong [2001], Roberts and Stramer [2002], Gill and Casella [2004]), hamiltonian Monte Carlo (Duane et al. [1987]), modified Metropolis-Hastings acceptance probabilities (Liu et al. [2000], Mira [2001], Green and Mira [2001]), Langevin-driven Metropolis-Hastings (Roberts and Tweedie [1996a]), the interaction of multiple chains

(Hukushima and Nemoto [1996], Hansmann [1997], Rosenthal [2000], Iba [2001], Myers and Laskey [2001], Zheng [2001], Corander et al. [2006], Jasra et al. [2007]), trans-dimensional algorithms (Green [1995], Liu and Sabatti [1998], Stephens [2000], Green and Mira [2001], Brooks et al. [2003], Cappé et al. [2003], Petris and Tardella [2003], Sission [2005]) and Monte Carlo variance reduction methods (McKeague and Wefelmeyer [2000], Mira and Sargent [2003]). More general approaches to improving mixing of traditional MCMC algorithms have focussed on the optimal scaling of proposal distributions (Roberts et al. [1997], Roberts and Rosenthal [1998a], Roberts and Rosenthal [2001], Neal and Roberts [2006], Neal et al. [2007], Bédard and Rosenthal [2008]) and on the construction of adaptive proposal distributions (Tierney and Mira [1991], Gilks and Wild [1992], Gelfand and Sahu [1994], Gilks et al. [1994], Gilks et al. [1995a], Gilks et al. [1998], Liu et al. [1997], Haario et al. [1999], Liu et al. [2001], Haario et al. [2001], Chauveau and Vandekerkhove [2002], Gasemyr [2003], Haario et al. [2005], Atchadé and Rosenthal [2005], Brockwell and Kadane [2005], Haario et al. [2006], Roberts and Rosenthal [2007], Roberts and Rosenthal [2008], Bai [2009], Bai et al. [2009]). These recent developments have greatly improved the empirical performance of MCMC algorithms, allowing for a substantial expansion of the domain of application of Bayesian methods, mainly by trading the simplicity of samplers for a reduction in Monte Carlo error. Partially as a result of this trend, the popularity of MCMC software such as WinBugs (Lunn et al. [2000]) or MrBayes (Huelsenbeck and Ronquist [2005]) has steadily increased.

In this paper we pursue an alternative strategy, by proposing a novel class of multi-chain samplers using standard MCMC algorithms as building blocks. These samplers do not employ temperatures as in Parallel Tempering

(PT) and they fully exploit cross-chain swap transitions to maximise mixing of one chain, which we call the "mother" chain. The drawbacks of tempered distributions are overcome by using different proposal distributions for each auxilliary chain rather than different marginal target densities. This strategy has two main advanteges. First, it allows mixing simulateneously over many proposal settings, which is an important advantage when analytical results on optimal proposal scaling or proposal adaptation are not easy to derive. A unique optimal proposal scaling might not in fact even exist as different proposal scalings might be optimal within different subsets of the domain unless regularity conditions on the target distribution are met. For instance, a single proposal kernel may not be optimal to explore both very narrow and wide peaks. Second, our multi-chain strategy can incorporate any combination of single-chain MCMC samplers. This is of great practical relevance allowing for fast implementation of our methods using existing computer code, especially in a research environment where distributed computing is becoming mainstream (Ren and Orkoulas [2007], Hu and Tsui [2008]).

In Section 1 we set out to motivate our approach by describing the framework of the cross-chains *swap* transitions common to PT, replica Monte-Carlo and Metropolis-coupled MCMC and we introduce the PHS algorithm. We show that mixing of the PHS mother chain is maximised at every iteration whereas its auxilliary chains are allowed to differ both in their proposal distributions and in their rates of mutual interaction. Convergence of the PHS transition kernel is proved. Waste-recycle and symmetrised versions of the PHS algorithm (sPHS) are illustrated. Section 2 compares the accuracy of single-chain Metropolis, PHS and of sPHS algorithm for sampling from Gaussian mixtures and for highly correlated 'banana-shaped' multivariate

densities (Haario et al. [1999], Haario et al. [2001]). These examples evaluate the relative accuracy of PHS respectively in case of multimodality and heavy tails. Sections 3 and 4 illustrate two applications of PHS for Bayesian model selection. First, we consider the standard problem of selecting significant main effects for the Gaussian linear regression model. Second, we use PHS to approximate marginal posterior inferences for the high level interactions defining the structure of a treed survival model. Section 5 concludes the paper discussing open problems and research opportunities in the field of multiple chains MCMC samplers.

1 Tempered and untempered multi-chains MCMC algorithms

In parallel to the development of novel single-chain samplers, the last twenty years have also witnessed the birth of multi-chain MCMC algorithms. The latter have been pioneered in conjunction with the use of tempered distributions by Swedensen and Wang [1987] and Hukushima and Nemoto [1996] in statistical mechanics and by Geyer [1991] in statistics. For each value of a chain index $m \in [1, M]$ with $M < \infty$ fixed, a tempered version of the target posterior distribution is defined by "powering up" its density

$$f_m(\theta \mid X) = \frac{f(\theta \mid X)^{\frac{1}{T_m}}}{C_m(X)}, \tag{1}$$

where $1 = T_1 \leq T_2 \leq ... \leq T_M < \infty$ is a vector of temperature levels and $C_m(X) = \int_\Theta f_m(\theta \mid X) d\theta$ is a positive and finite normalising constant depending on the temperature parameter T_m and on the data X. Here T_m acts as a smoother of the target distribution, so that the heated densities (1) have fatter tails and less pronounced modes than the target distribution

of interest $f_1(\theta \mid X)$. The key advantage of these algorithms is that detailed balance (DB) is preserved with respect to the marginal target distribution of each chain although different chains interact along the sampling process. Within-chain DB is achieved by coupling an *update* step with a *swap* step using the standard Metropolis rule. Metropolis-coupled MCMC (Geyer [1991], Hukushima and Nemoto [1996]), Parallel Tempering (Geyer and Thompson [1995]) and replica Monte Carlo (Swedensen and Wang [1987]) have been found to yield empirically reliable estimates especially when analogies to physical temperatures can be exploited to tune the sampler. In statistical mechanics, temperatures are chosen with reference to the physical properties of the systems being modeled, such as the energy barriers between electron excitation states implied by successive temperature levels. The equilibrium distributions sampled for applications in statistics seldom possess analogous interpretations, making temperature tuning a laborious process (Geyer and Thompson [1995], Neal [1996]). A second limitation to using auxilliary tempered distributions for Bayesian computations is that in general it is difficult to check whether a tempered posterior density is still proper. For instance, a sufficient condition ensuring $C_m(X) < \infty$ is that the Kullback-Leibler divergence $KL(f_1, f_m) = \int_\Theta f_1(\theta \mid X) \log \frac{f_1(\theta|X)}{f_m(\theta|X)} d\theta$ of the untempered proper posterior density $f_1(\theta \mid X)$ from its tempered version $f_m(\theta \mid X)$ is finite. When this is not the case, the Metropolis rule cannot be applied meaningfully neither for within-chain updates nor for cross-chains swaps. Even when tempered distributions are proper, recent developments show that when their modes tend to be very narrow, no matter how high a temperature is used mixing is always torpid (Woodard et al. [2009]). A third limitation of these multiple-chains algorithms is that posterior estimates can be calculated using all the M chains only if the samples arising from the

tempered chains are appropriately reweighted.

In what follows we let the indicator $s_i = 0$ if each of the M chains is updated independently at iteration i of the sampler and $s_i = 1$ if *swap* is chosen instead. The proposal probability $q_s^{'}(s_i \mid s_{i-1})$ describes how the two steps are combined along the sampling. For instance, Geyer [1991] adopts the deterministic proposal $q_s^{'}(s_i \mid s_{i-1}) = 1_{\{s_{i-1}=0\}}$, whereas Liu [2001] defines an independent PT sampler using $q_s^{'}(s_i \mid s_{i-1}) = s$ where $s \in (0,1)$ is a fixed swap proposal rate. Let the indexes j and k range over the set of chains $(1, ..., M)$ and let $\theta_{i,j}$ indicate the state of chain j at iteration i. We denote with $q_s^{''}(\theta_{i,j}, \theta_{i,k})$ the probability that, at iteration i, a swap is proposed between the current values of the chains with indexes (j,k). In Geyer [1991], Hukushima and Nemoto [1996] and Liu [2001] this proposal is taken to be uniform over all values of the ordered couple (j,k) with $k \neq j$ and. A swap is accepted with Metropolis probability

$$\alpha_s([\theta_{i,j}, \theta_{i,k}], [\theta_{i,k}, \theta_{i,j}]) = 1 \wedge \frac{f_j(\theta_{i,k}|X) f_k(\theta_{i,j}|X)}{f_j(\theta_{i,j}|X) f_k(\theta_{i,k}|X)}, \tag{2}$$

ensuring the reversibility of the sampler with respect to its joint target density $\mu(\theta_M \mid X, T_1, ..., T_M) = \prod_{m=1}^{M} f_m(\theta \mid X)$, where θ_M is the M-fold product of the random variable θ. When the independent updates of chain m are carried out using a single Metropolis-Hastings (MH) step with common proposal $q(\cdot)$, the joint transition kernel of the PT sampler is

$$K_{PT}(\theta_{M,i}, \theta_{M,i+1}) = (1 - q_s^{'}(s_i \mid s_{i-1})) \prod_{m=1}^{M} q(\theta_{i+1,m} \mid \theta_{i,m}) \alpha_{MH}(\theta_{i,m}, \theta_{i+1,m}) +$$
$$+ q_s^{'}(s_i \mid s_{i-1}) \sum_{j_i=1}^{M} \sum_{\substack{k_i=1 \\ k_i \neq j_i}}^{M} q_s^{''}(\theta_{j_i}, \theta_{k_i}) \alpha_s([\theta_{i,j}, \theta_{i,k}], [\theta_{i,k}, \theta_{i,j}]), \tag{3}$$

where $\theta_{M,i} = [\theta_{i1}, ..., \theta_{iM}]$ represents the joint state of all M chains at iteration i and we assume without loss of generality that $\int_\Theta \alpha_{MH}(\theta_{i,j}, \theta) q(\theta \mid$

$\theta_{i,j})d\theta = 1$. From (3) it can be seen that, with respect to the standard MH algorithm, PT can increase mixing for all chains through their successful swaps. Analogously to the MH algorithm, the irreducibility and aperiodicity of (3) critically depend on the proposal distributions for within-chains updates $q(\cdot)$ and on that of the cross-chains swaps $q_s^{''}(\cdot)$. A proof of convergence of the PT algorithm is sketched in Hukushima and Nemoto [1996].

1.1 Parallel Hierarchical Sampling

In this paper we consider an alternative class of multiple-chains MCMC samplers which proceed by carrying out both the following two steps at each iteration:

i) draw the index $m_i \in [2, ..., M]$ from a discrete proposal distribution $q_s^{''}(m_i \mid m_{i-1})$ and swap the current value of chain m_i with that of the mother chain;

ii) update independently the remaining $M - 2$ chains each having the same marginal target distribution $f(\theta \mid X)$.

At point i) above, we let $q_s^{''}(\cdot)$ be the swap proposal to emphasize the analogy with the PT algorithm. We label this class of algorithms parallel hierarchical samplers (PHS) because the mother chain is given a prominent role and the update of all chains is carried out in parallel analogously to PT. In particular the swap step in PHS always involves the mother chain, which represents the upper level in a hierarchy which lower level is composed of an array of auxilliary chains. To provide a simple proof of the reversibility of the PHS joint kernel, we assume that chains $(2, ..., M)$ are updated using a single MH step and that the transition kernels for these updates satisfy

the conditions illustrated in Tierney [1994] so that they are irreducible and aperiodic with respect to their common marginal target distribution $f(\theta \mid X)$. In addition, we assume that the symmetric proposal distribution $q_s^{''}(\cdot)$ allows for swaps between the mother chain and any of the other chains. Under these conditions the marginal transition kernel for the mother chain of the PHS algorithm is irreducible and aperiodic with respect to its target distribution. If the joint PHS transition kernel is also reversible with respect to the product density $\mu(\theta_M \mid X)$ having all marginals equal to $f(\theta \mid X)$, then $\mu(\theta_M \mid X)$ is its unique joint stationary distribution. The reversibility of the PHS is proved in the following theorem.

Theorem 1 *The PHS transition kernel satisfies detailed balance with respect to the joint distribution having product density or probability mass function* $\mu(\theta_M \mid X)$.

Proof The DB condition for the PHS algorithm is

$$\frac{\mu(\theta_{M,i})}{\mu(\theta_{M,i+1})} = \frac{K_{PHS}(\theta_{M,i+1}, \theta_{M,i})}{K_{PHS}(\theta_{M,i}, \theta_{M,i+1})}, \tag{4}$$

where $K_{PHS}(\theta_{M,i+1}, \theta_{M,i})$ is the PHS joint transition kernel. When the independent updates of the auxilliary chains are carried out via a MH step, the PHS joint transition kernel is

$$\begin{aligned} K_{PHS}(\theta_{M,i}, \theta_{M,i+1}) \quad = \quad & \sum_{m_{i+1}=2}^{M} q_s^{''}(m_{i+1} \mid m_i) \times \\ & \prod_{\substack{j=2 \\ j \neq m_{i+1}}}^{M} q_j(\theta_{i+1,j} \mid \theta_{i,j}) \alpha_{MH}(\theta_{i,j}, \theta_{i+1,j}), \end{aligned} \tag{5}$$

where the within-chain proposal probabilities $q_j(\cdot)$ and, as a consequence, the MH acceptance probabilities $\alpha_{MH}(\theta_{i,j}, \theta_{i+1,j})$ are now dependent on the chain index j. For each chain here we assume without loss of generality

that $\int_\theta \alpha_{MH}(\theta_{i,j}, \theta) q_j(\theta \mid \theta_{i,j}) d\theta = 1$. Each summand in (5) is the product of the marginal transition kernel for the swap transition and those of the $(M-2)$ independent MH updates for the remaining chains. The former coincides with the proposal $q_s^{''}(m_i \mid m_{i-1})$ because the PHS swap acceptance ratio is equal to one. This is a straightforward simplification of (2) when all temperatures are equal to 1. Under (5) the DB condition (4) can be rewritten as

$$\sum_{m_i=2}^{M} q_s^{''}(m_i \mid m_{i-1}) \prod_{\substack{j=2 \\ j \neq m_i}}^{M} f(\theta_{i,j} \mid X) q_j(\theta_{i+1,j} \mid \theta_{i,j}) \alpha_{MH}(\theta_{i,j}, \theta_{i+1,j}) =$$

$$= \sum_{m_i=2}^{M} q_s^{''}(m_{i-1} \mid m_i) \prod_{\substack{j=2 \\ j \neq m_{i-1}}}^{M} f(\theta_{i+1,j} \mid X) q_j(\theta_{i,j} \mid \theta_{i+1,j}) \alpha_{MH}(\theta_{i+1,j}, \theta_{i,j}). \quad (6)$$

For any given value of m_i, by the reversibility of the MH transition kernels with respect to the marginal density $f(\theta \mid X)$, the $M-2$ transition probabilities on the left-hand side of (6) are equal to their corresponding terms on the right-hand side. By taking $q_s^{''}(\cdot)$ symmetric with respect to m_i and m_{i-1}, for all values of m_i each summand on the left-hand side of (6) equals its corresponding term on the right-hand side, so that the equality (6) holds.

As can be seen from equation (5), the acceptance probability of the mother chain is one since this chain is implementing an independence MH algorithm using, as proposals, the MCMC samples accepted by a randomly chosen auxilliary chain. When the auxiliary chains are in stationarity, these proposals are indeed samples from the target and thus their acceptance probability is one. In other words, the mother chain is implementing exact Monte Carlo sampling of its marginal target distribution. From a different perspective, by picking its values from the current states of the auxilliary chains, the PHS

mother chain destroys the autocorrelation structure which is typically implied by Metropolis-Hastings type of algorithms. We emphasize this point in the next section by comparing the empirical autocorrelation functions (ACF) and the integrated auto-correlation time (IAT) of the chains generated by the PHS and the MH.

Equation (5) also shows that, as for the MH and PT algorithms, PHS does not require knowledge of the finite normalising constant of its marginal target distributions $C(X) = \int_{\Theta} f(\theta \mid X)$ and this makes it a suitable sampler for Bayesian applications. Furthermore, in light of the specific form of the joint PHS target density $\mu(\theta_M \mid X)$ the proposal distribution for within-chains updates in (5) can be generalised to $q_j(\theta_{i+1,j} \mid \theta_{i,-j})$, where $\theta_{i,-j} = (\theta_{i,1}, ..., \theta_{i,j-1}, \theta_{i,j+1}, ..., \theta_{i,M})$. This allows introducing mutual repulsion among the values proposed for the update of different chains along the lines of Green and Han [1991]. For example, when the set of conditional within-chain proposals $q_j(\cdot)$ are Gaussian, they can be constructed so that the joint proposal for the update for chains $(2, ..., M)$ is multivariate Normal with negative correlations.

Equations (3) and (5) show that the PHS swap proposal has a simpler form than that of PT. This is because at each iteration the PHS transition kernel mixes both update and swap steps whereas in PT they are alternated according to the proposal probability $q_s^{'}(s_i \mid s_{i-1})$. Thus, unlike PT, the PHS algorithm does not create unnecessary competition between local and global mixing when the update steps generate local transitions and the swaps produce larger jumps.

Whilst Theorem 1 proves reversibility of the joint PHS kernel, stationarity of each auxilliary chain with respect to their marginal target can be proved directly using each factor of the innermost product in (5). This is

shown in the Corollary below.

Corollary 1 *Each of the auxilliary chains of the PHS algorithm having transition kernel (5) is stationary with respect to the distribution* $f(\theta \mid X)$.

Proof For any value $\theta^* \in \Theta$, the stationarity condition for the auxilliary chain $j = 2, ..., M+1$ is

$$\int_{\Theta} f(\theta \mid X) K_{PHS,j}(\theta, \theta^*) d\theta = f(\theta^* \mid X), \tag{7}$$

where the transition kernel of the j-th chain is

$$\begin{aligned} K_{PHS,j}(\theta, \theta^*) = {} & q_s^{''}(j \mid m) f(\theta^* \mid X) + \\ & (1 - q_s^{''}(j \mid m)) q_j(\theta^* \mid \theta) \alpha_{MH}(\theta, \theta^*). \end{aligned} \tag{8}$$

Substituting (8) in (7) and using

$$\int_{\Theta} f(\theta \mid X) q_j(\theta^* \mid \theta) \alpha_{MH}(\theta, \theta^*) d\theta = f(\theta^* \mid X),$$

the identity of the left and right terms of equation (7) is verified.

In conjunction with the assumed aperiodicity condition and with the reversibility of the PHS kernel, this Corollary implies that the distribution of interest, $f(\theta \mid X)$, is the unique stationary distribution of any of the auxilliary PHS chains. Since the mother chain performs exact Monte Carlo sampling drawing from the auxilliary chains, (7) implies that the same target is the unique stationary distribution of the mother chain.

The PHS algorithm illustrated at points *i*) and *ii*) in this section and having transition kernel (5) represents one particular algorithm within a large class of parallel samplers which may differ for their within-chain updating rules and for their swapping rules. In what follows, we focus on comparing

the performance of this PHS algorithm with that of its symmetrised version (sPHS). The latter differs from the above definition of PHS in that at each iteration two chains are chosen uniformly at random and their current states are swapped, whereas all other chains are updated independently. With this symmetrised sampler, the mother chain loses its prominent role as the clearing house for swaps between any other pair of chains. Proof of convergence of the sPHS transition kernel can be derived along the lines of Theorem 1 above and it is not reported here.

2 Illustrative examples: multimodality and heavy tails

In this section we compare the empirical accuracy of single-chain MH versus PHS algorithms for generating samples from multimodal mixtures of Gaussian densities and from the heavy-tailed multivariate distributions of Haario et al. [1999] and Haario et al. [2001]. These two examples have been carefully constructed so as to compare the empirical accuracy of single-chain and multi-chain samplers having fixed their common computational cost.

2.1 MCMC samples from multimodal Gaussian mixtures

Due to their exponential tail behaviour, Gaussian mixtures provide a well-suited scenario for evaluating a sampler's ability to explore multimodal landscapes when the troughs between different modes are deep and far apart. In this example we consider bivariate Gaussian mixtures so as to be able to represent graphically their 2-dimensional probability contours. The number of components of the target mixture used in this section was set to 10, their

bivariate means μ_k for $k = 1, ..., 10$ were generated uniformly at random over the square $(-10, 10)^2$ and their variances were all set to 1. The covariance between the first and second dimension of each mixture component is $\Sigma_k(1,2) = \Sigma_k(2,1) = \frac{k}{M+1} sign(u \leq 0.5)$ where $u \sim Uniform(0,1)$ and $k = 1, ..., 10$, so that the two dimensions of the successive mixture components are progressively more correlated, either positively or negatively. The 10 mixture weights w_k were generated uniformly at random and then normalised so as to sum to 1. Figure 1 shows the probability contour of this Gaussian mixture density. As expected, different bivariate modes are separated by throughs where the density is numerically zero.

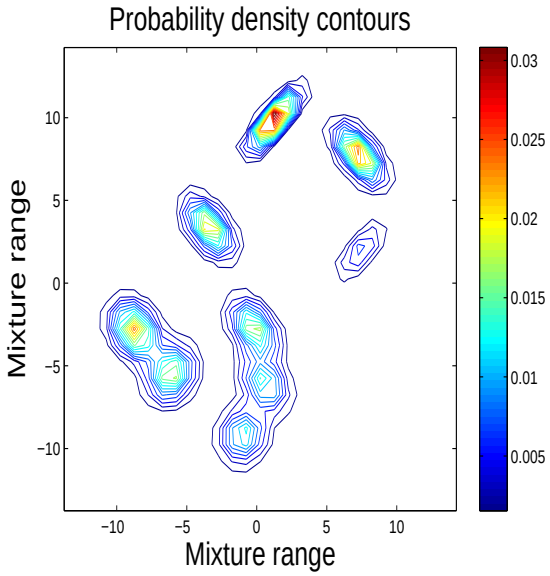

Figure 1: probability contour of the 10-components bivariate Gaussian mixture target density. Sampling from this distribution by single chain random walk algorithms is made difficult by the deep throughs between different modes where the target density is almost zero.

To generate samples from this Gaussian mixture via MCMC, we first use a standard random walk MH algorithm whith proposal distribution $q(\theta_i \mid \delta)$ taken to be Normal with mean equal to the current state θ_i and variance δ. Here we let $\delta = 3$, ensuring an average acceptance rate close to the optimal 0.234 (Roberts and Rosenthal [2001]).

As a first alternative to this vanilla algorithm we consider an other Metropolis-Hastings sampler using a version of the above proposal enriched by a Langenvin term (Roberts and Tweedie [1996a]). At each iteration i this random walk proposal is Normal with mean $\theta_i + \nabla \log f(\theta_i \mid w_{1:10}, \mu_{1:10}, \Sigma_{1:10})$ and variance δ, where $\log f(\theta_i \mid w_{1:10}, \mu_{1:10}, \Sigma_{1:10})$ is the logarithm of the probability density for the Gaussian mixture and ∇ stands for the matrix of its partial derivatives with respect to θ. Using the same random walk proposal distribution with $\delta = 3$ we attain an acceptance rate comparable to that of the plain MH sampler.

We compare the accuracy of these two single-chain samplers with that of PHS as described in the previous section and with a symmetrised version of PHS (sPHS). The latter differs from the above definition of PHS in that at each iteration two chains are chosen uniformly at random and their current states are swapped, whereas all other chains are updated independently. With this symmetrised sampler, the mother chain loses its prominent role as the clearing house for swaps between any other pair of chains. For the within-chain updates of both parallel algorithms we use the same type of random walk proposal $q_m(\theta_{i,j} \mid \delta_j)$ as for the MH algorithm but we let the variance δ_j be chain-specific. In what follows we let $\delta_j = \frac{j}{4}$ for $j = 2, ..., M$ and $M = 20$, so as to cover from relatively small to large proposal variances, including that of the vanilla random walk MH sampler. This range of values for the spread of the within-chain proposal distributions leads to acceptance

rates within the range $(0.15, 0.76)$.

We calculate Monte Carlo estimates of the two-dimensional Gaussian mixture mean using respectively the samples generated by the MH algorithm, the mother chain of the PHS algorithm and the samples generated by all chains of the sPHS algorithm. In a waste-recycle perspective, pooling of the sPHS samples is implemented using two different weighting schemes. In the first scheme we calculate the naïve average of the Monte Carlo estimate of each chain

$$\hat{\theta}_N = \frac{\sum_{j=1}^{M} \hat{\theta}_j}{M}, \tag{9}$$

where $\hat{\theta}_j$ is the empirical two-dimensional mean vector calculated from chain j of the sPHS sampler. In the second scheme, we calculate the weighted Monte carlo estimate

$$\hat{\theta}_{S,i} = \sum_{j=1}^{M} w_{i,j} \hat{\theta}_{i,j}, \tag{10}$$

where $\hat{\theta}_{i,j}$ is the average of chain j for the mean component $i = 1, 2$ and the weight associated to this component of the mean estimate of chain j is

$$w_{i,j} = \frac{\frac{1}{IAT(i,j)}}{\sum_{j=1}^{M} \frac{1}{IAT(i,j)}}, \tag{11}$$

and $IAT(i, j)$ is the integrated auto-correlation time of component i of chain j (Sokal [1996]). The latter is estimated using the Gamma method of Wolff [2004]. The pooled estimator $\hat{\theta}_S$ relatively downweights estimates of the mean vector associated to the poorly mixing chains. This weighting scheme is useful to reduce the Monte Carlo error of the pooled mean estimator when the proposals chosen for the within-chains updates produce substantially different mixing behaviours.

We compare the empirical accuracy of the three algorithms by repeating these sampling processes for 100 independent runs. In order to make the computational cost for all samplers comparable, each repetition the PHS and sPHS algorithms was run 5000 iterations whereas those of the MH algorithm (random walk and Langevin) were run for $5000 \times M$ iterations using $M = 20$ auxilliary chains. As a measure of accuracy of a sampler we use its empirical mean square error (MSE) about the mean

$$\text{MSE} = \frac{\sum_{j=1}^{100} (\theta_T - \hat{E}^j(\theta))(\theta_T - \hat{E}^j(\theta))'}{100}, \tag{12}$$

where θ_T is the true value of the 2-dimensional Gaussian mixture mean and $\hat{E}^j(\theta)$ is the estimate of θ derived from the jth repetition of the sampling process obtained from the jth repetition of the sampling process using either of the random walk MH algorithm (rwMH), the random walk Langevin MH algorithm (rwLMH), the PHS mother chain (PHSm), the pooled PHS estimators using the the naïve (PHSN) and Sokal (PHSS) weights and the pooled sPHS estimators (sPHSN, sPHSS).

Table 1 reports the MSE estimates arising from this simulation experiment. Having fixed a common computational cost, the accuracy of the mean estimates of either of the multiple-chains samplers is found to be superior at least by one order of magnitude with respect to that of the estimates obtained by the MH algorithms. Table 1 also shows that the symmetrised parallel sampler sPHS (last two rows) is in this case slightly less accurate than its asymmetric version PHS (third to fifth rows). Finally, the empirical accuracy of the weighted estimators PHSN and PHSS is found to be comparable to those of the PHS mother chain. This is a somewhat surprising result, since the weighted estimates are derived using 20 times as many MCMC samples. The equivalence of the empirical MSEs of the three

estimators uggests that the PHS mother chain by construction efficiently incorporates all the information included in all samples of its auxilliary chains. The MSE discrepancies shown in Table 1 are essentially due to the fact that

Sampler	**MSE**
rwMH	9.54
rwLMH	12.61
PHSm	0.75
PHSN	0.77
PHSS	0.78
sPHSN	0.94
sPHSS	0.91

Table 1: empirical mean squared errors of the MCMC estimator for the two-dimensional Gaussian mixture mean. The accuracy of the multiple-chains samplers is found to be superior to that of the estimates obtained by the MH algorithms having fixed a common computational cost.

the draws included in the PHS mother chain are far less correlated than those of the single chain algorithms. This point is illustrated in Figure 2, which shows the empirical autocorrelation functions (ACF) of one of the generated MH chains, of one PHS mother chain and of the auxilliary chain having the same proposal spread as that of the MH algorithm. Finally, Table 1 shows that the symmetrised parallel sampler sPHS (last two rows) is in this case slightly less accurate than its asymmetric version PHS (third to fifth rows).

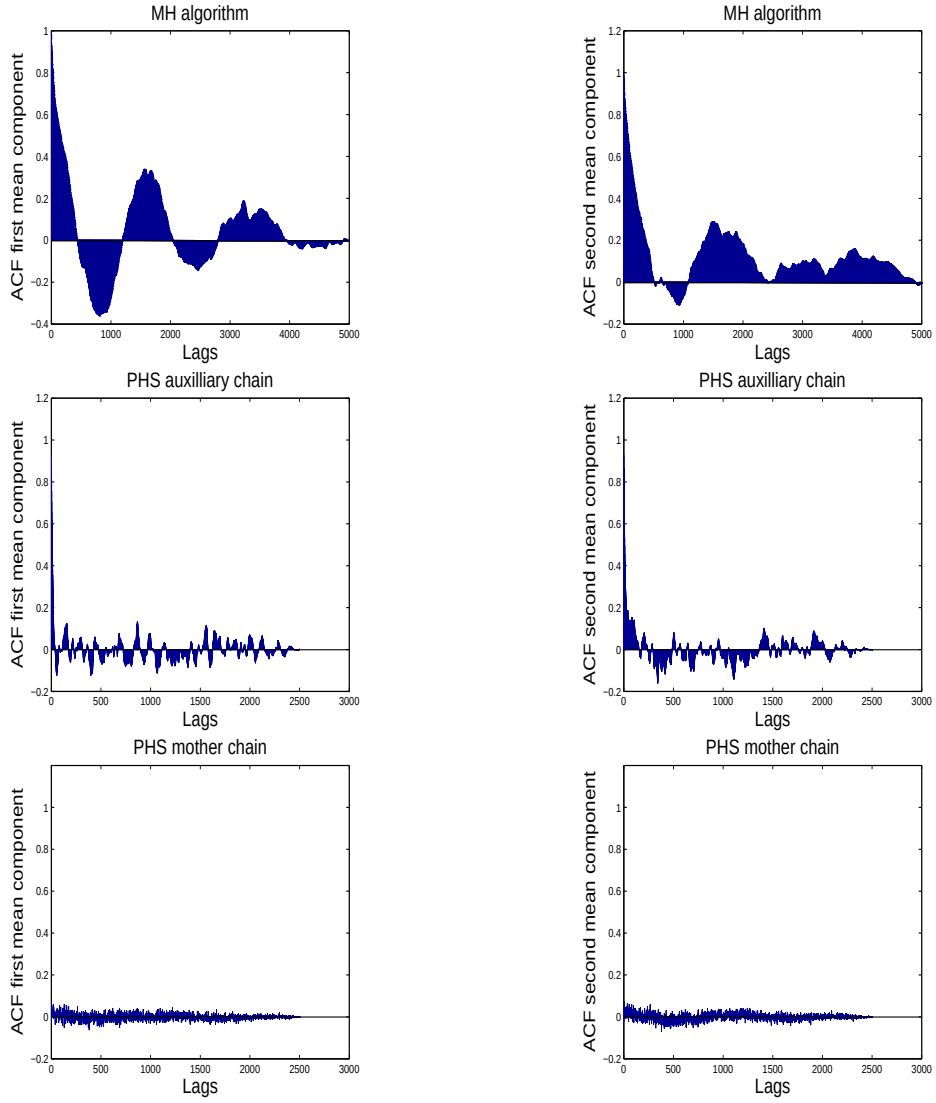

Figure 2: empirical ACF for chains sampling the posterior distribution of each component of the bivariate Gaussian mixture mean. The serial dependence generated by the random walk MH algorithm (first row) and to a lesser extent by the PHS auxilliary chain (second row) is contrasted with the lack of correlation of the PHS mother chain (last row). ation of the PHS mother chain (last row). These differences are mirrored by the corresponding IAT, which are $(177.97, 143.05)$ for MH, $(17.10, 13.14)$ for the auxilliary PHS chain and $(0.48, 0.47)$ for the PHS mother chain.

Figure 3 illustrates the empirical relationship between the accuracy of the PHS mother chain estimates and the number of auxilliary chains for this Gaussian mixture example. The MSE of the mean estimator decreases until the number of auxilliary chains reaches the number of Gaussian components in the target density, whereas for $M > 10$ the sampler achieves almost no further gains in accuracy. These results suggest that, when a target probability density is multimodal, multiple-chain samplers where the number of auxilliary chains roughly matches the number of modes of their target can yield much more accurate Monte Carlo estimates with respect to single-chain samplers having the same computational cost.

2.2 MCMC samples from heavy-tailed Gaussian mixtures

In this section we compare the empirical MSEs (12) for the MH, PHS and sPHS algorithms when their multidimensional target distribution is the non-linear transformation of a Gaussian distribution of Haario et al. [1999]. As opposed to the previous example, here we focus on evaluating the samplers' accuracy when their target density is heavy-tailed. Let X be a d-dimensional Gaussian random variable with diagonal covariance matrix C having all non-zero entries equal to one but for its upper left entry, which is set to 100. The probability density of its non-linear transformation

$$Y = [X_1, X_2 + b(X_1^2 - 100), X_3, ..., X_d],$$

is such that the contours of its first two dimensions are twisted and elongated in a shape resembling that of a symmetric banana. Its non-linearity increase with the value of the hyper-parameter b. In this section we use $d = 8$ and

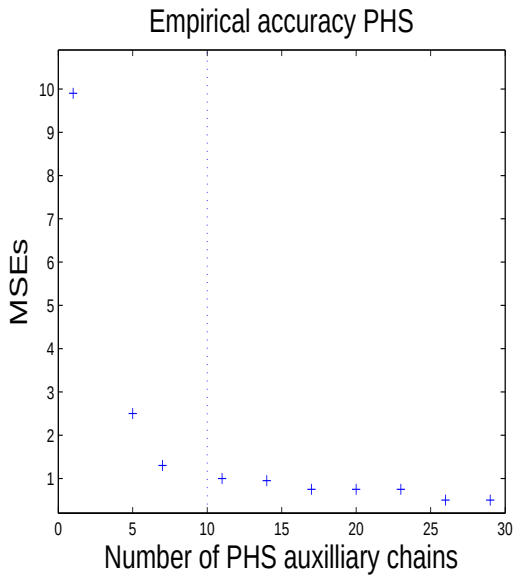

Figure 3: empirical relationship between accuracy of the PHS mother chain estimates (MSE) and number of auxilliary chains for this Gaussian mixture example. The decrease in the MSE of the mean estimator suggests that multiple-chain samplers where the number of auxilliary chains roughly matches the number of modes of their target can yield much more accurate Monte Carlo estimates with respect to single-chain samplers having the same computational cost.

$b = 0.03$ so that our numerical results can be compared directly with Haario et al. [1999] and Haario et al. [2001].

A comparative evaluation of the empirical precision for MH, PHS and sPHS was carried out by estimating the MSEs of their respective estimators for the 8-dimensional mean, as in section 2.1. Each sampler was run 100 independent times using the same Gaussian proposal distributions and the same number of iterations as in the previous example. The spread of the uniform random walk proposal distribution for all samplers was set at 1, yielding a MH acceptance ratio of approximately 0.4. The starting point for all samplers was set at the origin of the target support $\mathcal{R}^8$. The left panel in Figure 4 shows the 5000 samples of the PHS mother chain for the first and second dimensions of the banana-shaped target distribution, closely matching the theoretical contours represented in Haario et al. [1999]. Table 2 reports the empirical MSEs for the mean estimators of the random walk MH algorithm (rwMH), the adaptive random walk MH algorithm (arwMH of Haario et al. [1999]), of the PHS mother chain (PHSm), the PHS naïve weighted mean estimator (PHSN), PHS Sokal weighted estimator (PHSS), the sPHS naïve weighted estimator (sPHSN) and the sPHS Sokal weighted estimator (sPHSS). As in the previous example, having controlled for computational cost, the two versions of the parallel sampler using 20 auxilliary chains yield mean estimators of comparably better precision with respect to those of the random-walk MH algorithm. The last two rows show that, unlike for the multi-modal Gaussian mixture example the symmetrised sampler sPHS is found to be slightly more accurate than PHS (third to fifth rows). Table 2 also shows that the accuracy of the single-chain adaptive MH sampler is much better than for the random-walk MH, its empirical MSE being of the same order of magnitude than that of the multiple-chains samplers. Finally,

as in Table 1, the MSEs of the weighted estimators PHSN and PHSS are comparable to those derived from the PHs mother chain. The right panel in

Sampler	**MSE**
rwMH	23.82
rwaMH	5.22
PHSm	4.89
PHSN	4.77
PHSS	4.92
sPHSN	3.15
sPHSS	3.15

Table 2: empirical mean squared errors of the MCMC estimator for the eight-dimensional banana-shaped distribution. The two versions of the parallel sampler using 20 auxilliary chains yield mean estimators of comparably better precision with respect to those of the random-walk MH algorithm. The precision of the multiple chains samplers is found to be comparable to that of the single-chain adaptive MH sampler.

Figure 4 illustrates the empirical relationship between the number of auxilliary chains and MSE of the mean PHSm estimator. As for the case of multimodality, when the target density is heavy-tailed parallelisation yields more accurate MCMC estimators of the mean although the successive gains in precision decrease as the number of chains grows.

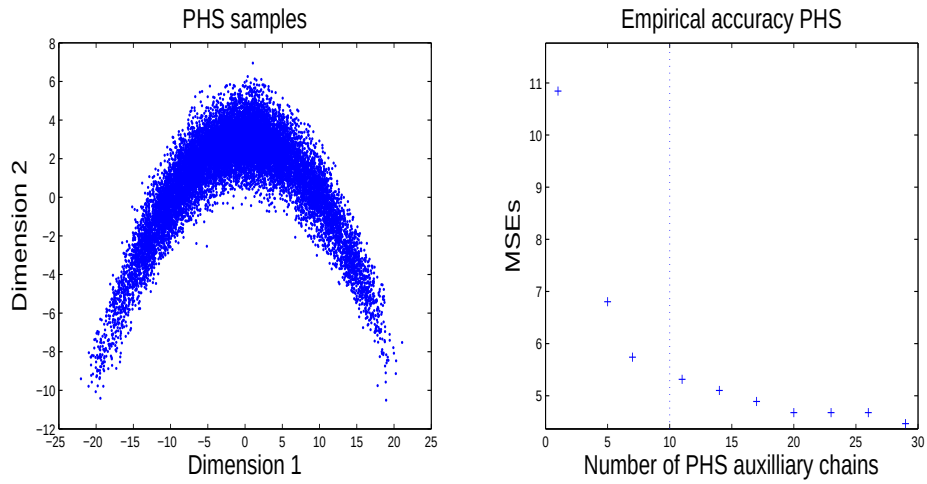

Figure 4: on the left, 5000 samples of the PHS mother chain for the first and second dimensions of the symmetric banana-shaped target distribution, closely matching the theoretical contours represented in Haario et al. [1999]. On the right, empirical relationship between the number of auxilliary chains and MSE of the mean PHSm estimator. for this heavy-tailed target density parallelisation yields more accurate MCMC estimators of the mean although the successive gains in precision decrease as the number of chains grows.

3 Application to the selection of covariates for the Bayesian linear regression model

The selection of main effects for the Bayesian Gaussian linear regression has been addressed using MCMC methods by Mitchell and Beauchamp [1988], Smith and Kohn [1996], George and McCulloch [1993], Carlin and Chib [1995], George and McCulloch [1997], Raftery et al. [1997], Kuo and Mallick [1998], Dellaportas et al. [2002] and Clyde and George [2004] among many others.

Using the same notation as in George and McCulloch [1997], we let the distribution of the n-dimensional random vector Y be multivariate Gaussian with mean $X_\gamma \beta_\gamma$ and covariance matrix $\sigma^2 I_n$, being (σ, β, γ) a priori unknown. The p-dimensional model index γ has elements γ_j taking value one if the jth covariate is used for the computation of the mean of Y and zero otherwise. The binary vector γ can thus take 2^p distinct values. Here β_γ and X_γ include respectively the elements of the p-dimensional column vector β associated to the statistically significant components of γ and the corresponding columns of X. The latter is a $n \times p$ matrix representing p potential predictors for the mean of Y. Within this framework, the variable selection problem is adressed in the current Bayesian literature using the marginal model inclusion probabilities $P(\gamma_j \mid Y, X)$ where $j = 1, ..., p$. These measure the marginal fitness of each covariate to explain the outcome data Y using the assumed linear model structure. When the number of potential predictors is large, model inclusion probabilities can be used to select a smaller number of covariates to focus the modeling effort. As such, these probabilities can be seen as useful descriptors of the marginal linear relationships between each covariate and the outcome variable.

When the model space is too large for implementing exhaustive search algorithms, model inclusion probabilities can be approximated via MCMC by sampling from the model space. In this section we evaluate the reliability of the MH sampler versus the PHS for generating draws from the marginal posterior probability of the model index,

$$P(\gamma \mid Y, X) \propto P(\gamma)P(Y \mid \gamma, X).$$

Here we adopt the same form of the marginal posterior probability of γ as in Nott and Green [2004], that is

$$P(\gamma \mid Y, X) \propto (1+n)^{-\frac{S(\gamma)}{2}} \left(Y'Y - \frac{n}{n+1} Y' X_\gamma (X_\gamma' X_\gamma)^{-1} X_\gamma' Y \right)^{-\frac{n}{2}}, \quad (13)$$

where $S(\gamma) = \sum_{j=1}^{p} \gamma_j$ As noted by George and McCulloch [1997], Denison et al. [1998] and Nott and Green [2004], efficient MCMC simulation from the above marginal probability mass function is hampered by the sheer dimension of the model space and by the presence of collinearity among the p model dimensions (Smith and Kohn [1996]). In particular, when collinearity is sufficiently strong or when the sample size is less than the number of covariates p, the target distribution (13) can be highly multimodal. In this situation the example in section 2 suggests that single-chains samplers can yield unreliable results.

Here we compare the consistency of the MH and PHS samplers using a set of physiological measurements taken at sea level in preparation to a carefully designed research expedition to mount Everest (Grocott et al. [2008]). We focus on selecting significant predictors of the blood concentration of lactic acid at the anaerobic respiration threshold (LAT), which is related to endurance performance (Yoshida et al. [1987]). This study is motivated by the fact that, although the biological mechanisms leading to the production

of lactic acid in tissues are well characterised, the correlations between LAT and other metabolites in blood are less understood. The data reports the blood LAT along with the concentrations of 50 relevant metabolites for 171 subjects. The outcome variable LAT and its covariates were log-transformed prior to analysing the data using model (13). As shown in Figure 5, the complexity of this data does not arise from a very large number of covariates as in West [2003] but in the collinearity among the predictors, which exhibit correlations ranging from -0.98 up to 0.99.

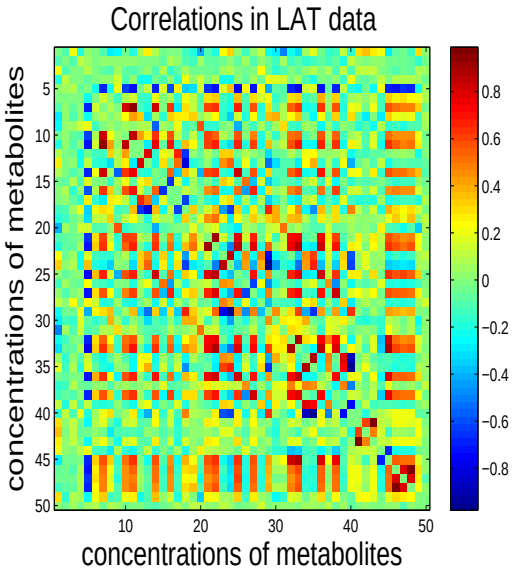

Figure 5: empirical correlations among the 50 LAT predictors. The intrinsic biological relationships among many of the metabolites result in a strong collinearity of their measured blood concentrations, ranging from -0.98 up to 0.99.

Consistency of the MCMC estimates for the marginal model inclusion probabilities can be assessed using their Monte Carlo standard errors (MCSEs).

As illustrated by Geweke [1992], Nott and Green [2004] and by George and McCulloch [1997], the MCSE for the inclusion probability of the jth predictor is

$$MCSE(\bar{\gamma}_j) = \sqrt{\frac{1}{N}\sum_{|h|<N}\left(1-\frac{|h|}{N}\right)A_j(h)},$$

where $\bar{\gamma}_j = \sum_{i=1}^{N}\gamma_j^i/N$, γ_j^i is the ith MCMC draw for the jth covariate and $A_j(h)$ is the lag h autocovariance of the chain of realisations for $\{\gamma_j\}_{j=1}^{N}$. For ergodic Markov chains, as $N \to \infty$ the MCSE converges, up to an additive constant independent of its transition kernel, to the MCMC standard error $\sigma_{g,K}$ (Mira and Geyer [1999]) where $g(\gamma_j) = E(\gamma_j \mid Y, X)$ for this example. Empirical MCSEs are calculated in this section using the empirical autocovariances of the chains representing inclusion or exclusion of each LAT covariate.

Independent batches of PHS and MH chains were run to estimate each covariate's model inclusion frequencies and their empirical MCSE. The former multi-chain sampler was run for twenty thousand iterations using fifty auxilliary chains and the length of the MH algorithm chains was set at one million iterations to match computational costs. Sampling was repeated twice so as to visually compare the consistency of the estimated model inclusion frequencies for each sampler. The prior inclusion probabilities were set at $P(\gamma_j = 1) = 0.5$ for $j = 1, ..., 50$ for both algorithms. The MH algorithm was implemented using an independent sampler proposal with inclusion probability 0.5 for each covariate. The same proposal was used for all within-chain updates of the PHS algorithm, whereas swaps between the current states of all chains were proposed uniformly at random. The top panels in Figure 6 compare the estimated model inclusion frequencies for the 50 LAT covariates respectively arising from the two independent runs of

the MH algorithm and of the PHS mother chain. The estimates of perfectly consistent samplers would be aligned exaclty on the 45 degrees dotted line in each plot. The correlation between the estimated inclusion frequencies obtained by the two runs of the samplers are respectively 0.27 for MH and 0.93 for the PHS mother chain, suggesting that PHS produces far more reliable inferences for the model inclusion probabilities with respect to MH in presence of strong collinearity. This conclusion is supported by the bottom-left panel in Figure 6, which shows that the ratio of the estimated MCSEs for the PHS and MH algorithms is consistently less than one. The bottom-right panel in Figure 6 represents the PHS estimated model inclusion frequencies for the 50 LAT covariates. Using the predictively optimal threshold of 0.5 inclusion probability (Barbieri and Berger [2004]) only two of the 50 covariates, that are the work rate and the respiratory exchange ratio, are found significant.

4 Application to the estimation of the structure of a survival CART model

In regression and classification trees (CART) a sample is clustered in disjoint sets called leaves. These are the final nodes of a single-rooted binary partition of the covariates space which is referred to as the tree structure. Within each leaf, the response variable is modeled consistently with the regression, or classification or with the survival analysis frameworks (Breiman et al. [1984]). As opposed to standard parametric regression methods, such as those entertained in section 4, CART trees are tailored to inferring in-

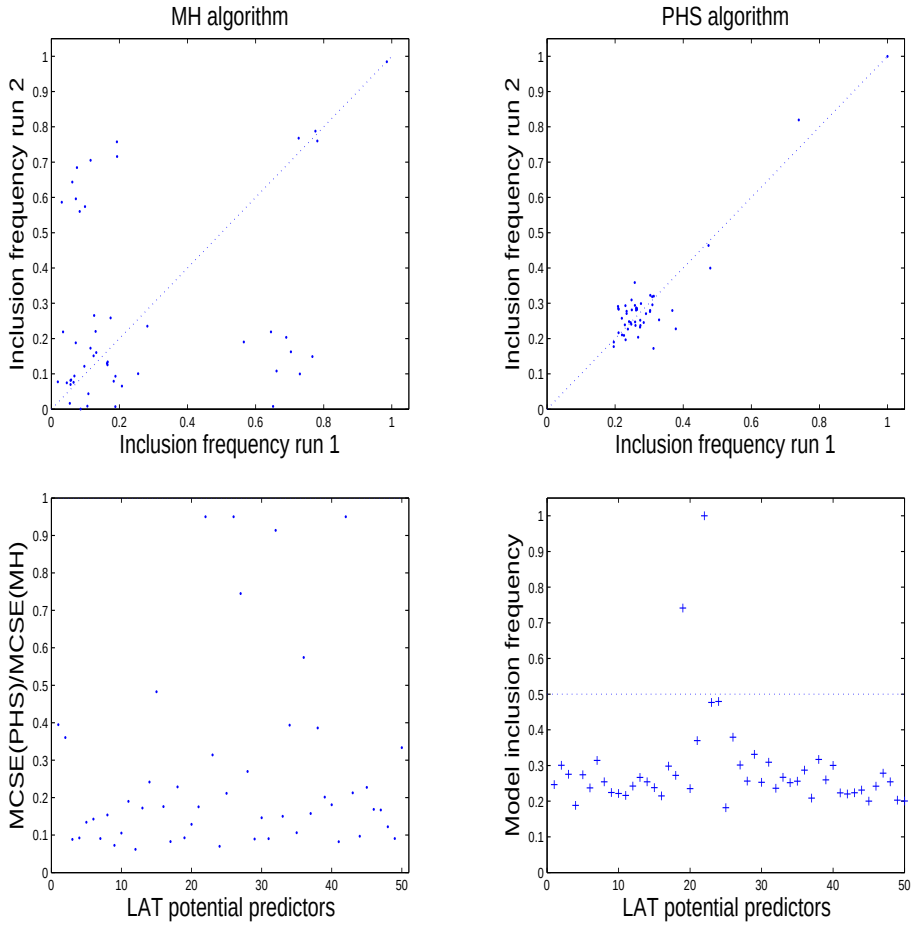

Figure 6: on top, estimated model inclusion frequencies for the 50 LAT covariates arising from two independent runs respectively of the MH algorithm and of the PHS mother chain. The correlation between the estimated inclusion frequencies obtained by the two runs of the samplers are respectively 0.27 for MH and 0.93 for the PHS mother chain, suggesting that the latter sampler produces far more reliable inferences for the model inclusion probabilities with respect to singel chain MH. This conclusion is supported by the bottom-left panel, which shows that the ratio of the Monte Carlo standard errors for the PHS and MH algorithms is consistently less than one. Bottom right: PHS estimated model inclusion frequencies for the 50 LAT covariates. Using the predictively optimal threshold of 0.5 only the work rate and the respiratory exchange ratio, are found as significant LAT predictors.

teractions among different covariates to fit the statistics of the response variable within different leaves. Tree models also form the basis of several non-parametric classification and regression methods, among which random forests (Ho [1998], Breiman [2001]), bagging (Breiman [1996]) and boosting (Breiman [2004]). Bayesian CART models appeared in the literature in Chipman et al. [1998] and Denison et al. [1998]. The MCMC model search algorithms developed in these two papers regard the tree structure as an unknown parameter and explore its marginal posterior distribution using the MH algorithm. Here we focus on tree models for randomly right-censored survival data (Gordon and Olshen [1995], Davis and Anderson [1989], M.Leblanch and J.Crowley [1992a], M.Leblanch and J.Crowley [1992b]). The first Bayesian survival tree model has been proposed by Pittman et al. [2004], who adopted a Weibull leaf sampling density and a step-wise greedy model search algorithm based on the evaluation of all the possible splits within each node. The main strength of this model search algorithm is that it quickly locates the most prominent posterior modes within large tree spaces. Its main weakness is that, when a large number of low posterior probability models yield predictions departing from those of the modal trees, predictive intervals based on maximum a posteriori (MAP) trees underestimate the uncertainty associated to future suvival events. The key difficulty encountered by single chain random walk MCMC tree search methods is in fact their limited ability to effectively explore such highly multimodal model spaces. From this perspective, a key advantage of multi-chain MCMC algorithms for tree model search is that they allow a variety of cross-chains transitions swapping features of the current state of different chains, such as tree branches or the covariate thresholds defining different data clusters. Therefore, thanks to their cross-chains transitions,

multi-chain samplers can only improve mixing in tree space with respect to their single-chain analogues. Here we implement a fully Bayesian analysis of the marginal tree posterior distribution using the PHS algorithm under an exponential leaf likelihood. The latter allows a closed-form evaluation of the tree marginal likelihood, which is a key requirement for implementing computationally efficient MCMC model search algorithms.

4.1 Tree structure marginal posterior distribution

Let the survival times $\{t_j\}_{j=1}^n$ be independent random variables conditionally on the tree structure and on the exponential leaf parameters. In what follows a tree structure will be represented by the couple (ℓ, ζ) where ℓ is the number of tree leaves and $\zeta = [\zeta_1, ..., \zeta_\ell]$ is a collection of disjoint subsets of the covariate space $\mathcal{X}$ corresponding to each of the leaves. Under the exponential likelihood, the joint sampling density of the survival times is

$$f(t \mid X, \delta, \ell, \zeta, \lambda_\zeta) = \prod_{k=1}^{\ell} \prod_{j=1}^{n} \left(\lambda_k^{\delta_j} e^{-\lambda_k t_j} \right)^{1_{\{X_j \in \zeta_k\}}}, \tag{14}$$

where $\lambda_\zeta = [\lambda_1, ..., \lambda_\ell]$ is the ℓ-dimensional vector of exponential parameters for each of the tree leaves and δ_j takes value 1 for exact observations and 0 for right censored observations. The indicator $1_{\{X_j \in \zeta_k\}}$ is 1 if the covariate profile of the jth subject is included in $\zeta_k \subseteq \mathcal{X}$ and 0 otherwise. Under a discrete uniform prior for the tree structure, the marginal posterior probability $P(\ell, \zeta \mid t, X, \delta)$ can be obtained, up to a multiplicative constant, by integrating (14) with respect to the conditional prior distribution for the array of leaf parameters λ_ζ. To derive a closed-form expression of the marginal tree likelihood we adopt an independent conjugate Gamma prior for each

leaf with probability density

$$P(\lambda_\zeta \mid \ell) = \prod_{k=1}^{\ell} \frac{b_k^{a_k}}{\Gamma(a_k)} \lambda_k^{a_k - 1} e^{-\lambda_k b_k}.$$

For this specification of the prior structure, the joint posterior of the tree structure is

$$P(\ell, \zeta \mid t, X, \delta) \propto \prod_{k=1}^{\ell} \frac{b_k^{a_k}}{\Gamma(a_k)} \frac{\Gamma(a + \sum_{j=1}^{n} \delta_j 1_{\{X_j \in \zeta_k\}})}{(b + \sum_{j=1}^{n} t_j 1_{\{X_j \in \zeta_k\}})^{a + \sum_{j=1}^{n} \delta_j 1_{\{X_j \in \zeta_k\}}}}. \quad (15)$$

A natural extension of model (15) yielding a closed-form likelihood function is to assume Weibull-distributed survival times as in Pittman et al. [2004] but in this case the marginal tree likelihood cannot be derived in closed form as no conjugate prior is available for the Weibull shape parameter. Approximate marginal tree likelihoods can be derived in this case using the Laplace or the Schwartz approximations. Here it can be noted that the penalty term of both these approximations increase with the number of tree leaves ℓ but for any fixed number of leaves the Schwartz approximation favours trees allocating the data more unevenly across leaves whereas the Laplace approximation does not favour unbalanced trees. In light of this difference, when the cluster sizes defined by the number of observations within each tree leaf are of interest, under a Weibull leaf likelihood we find recommendable adopting the Laplace approximation to the marginal posterior in conjunction with a suitable prior on the volume of the tree structure parameters $(\zeta_1, ..., \zeta_\ell)$.

4.2 Marginal posterior inference for the tree structure

The main challenge for constructing efficient within-chain proposal distributions for CART models is the lack of a natural distance metric between

different trees. This general issue has been noted, for instance, by Brooks et al. [2003] in the context of the reversible jump MCMC algorithm (Green [1995]). The within-chain proposal distribution used here generalizes the approaches of Denison et al. [1998] and Chipman et al. [1998] by devising two additional within-chain transitions. For the within-chain updates we propose a transition within the tree space using the following five moves:

1) Insert: sample a leaf at random and insert a new split by randomly selecting a new splitting rule.

2) Delete: sample at random a leaf pair with common parent and at most one child split and delete it.

3) Change: resample at random one splitting rule.

4) Permute: sample a random number of splits and permute at random their splitting rules.

5) Graft: sample at random one of the tree branches and graft it to one of the leaves of a different branch.

Chipman et al. [1998] noted that their MCMC algorithm can effectively resample the splitting rules of nodes close to the tree leaves but the rules defining splits close to the tree root are seldom replaced. In our specification of the within-chain transitions, move 4) aims at improving sampling of the splitting rules at all levels of the tree structure. Furthermore, the fifth move type allows the sampler to jump to a tree structure distinct from the current one without changing any of its splitting rules but only their combinations.

To take full advantage of our multiple-chains algorithm, we also devised two types of cross-chains transitions. The first is the cross-chains version of

the transitions of types 1), 3) and 5), swapping the elements of the tree structure required to perform corresponding pairs of transitions across chains. The second class of cross-chains transitions includes a whole tree swap between chains.

At iteration i, the PHS algorithm for this example proceeds as follows:

i) choose at random one of the auxilliary chains $m_i \in [2, M]$ and propose at random one of the two cross-chains moves, accepting the swap with probability 1.

ii) update each of the remaining $M-2$ chains independently using the five types of within-chain transitions and the stndard Metropoilis-Hastings acceptance probability.

4.3 Analysis of a set of cancer survival times

Colorectal adenocarcinoma ranks second as a cause of death due to cancer in the western world and liver metastasis is the main cause of death in patients with colorectal cancer (Pasetto et al. [2003]). In this section we analyse a set of 622 exact and right-censored survival times of patients with liver metastases from a colorectal primary tumor. The data were collected along with their clinical profiles by the International Association Against Cancer (`http://www.uicc.org`). Table 1 reports a description of the nine available clinical covariates. This survival data has been analyzed among others by Hermanek and Gall [1990] using non-parametric methods, by Antoniadis et al. [1999] using their wavelet-based method for estimating the survival density and the instantaneous hazard function and by Kottas [2006], who employed a Dirichlet process mixture of Weibull distributions to derive a Bayesian non-parametric estimate of the survival density and of the hazard

function. Haupt and Mansmann [1995] employed this dataset to illustrate the non-parametric tree fitting techniques for survival data implemented in the `S-plus` function *survcart*. This Section shows that the estimates of (b, ζ) obtained using the PHS algorithm and the approximate marginal posterior (15) discriminate statistically different survival groups based on differences among their covariate profiles. According to the latest EUROCARE de-

Symbol	**Description**	**Range**
DLM	Diameter largest LM	$(1, 20)$mm
AGE	Age	$(18, 88)$years
TD	Diagnosis of LM	synchrone/metachron with CPT
SEX	Gender	M = 55.8%, F = 44.2%
LI	Lobar involvement	unilobar/bilobar
NLM	Number of LM	$(1, 20)$
LRD	Locoregional disease	yes/no
TNM	Metastatic stage	local/regional/distant
LOC	Location PT	colon/rectum

Table 3: description of the covariates for the liver dataset. The data include several types of clinical covariates, such as continuous (DLM), discrete (AGE, NLM) and categorical (all others). This analysis aims at discriminating statistically different survival groups based on differences among their covariate profiles.

scriptive study, colorectal cancer survival rates at five years from surgery are consistently close to 50% for all the monitored European countries (Berrino et al. [2009], Sant et al. [2009]). We incorporate this information in the analysis of the present data by setting the Gamma prior hyper-parameters of the exponential survival rates within the tree leaves to $a_k = 9$ and $b_k = 0.1$ for

$k = 1, ..., \ell$. Using this informative prior, under the exponential likelihood we obtain a prior predictive median survival time of roughly 60 months, reflecting the descriptive statistic reported by the EUROCARE study.

A PHS using nine auxilliary chains was run for two hundred thousand iterations, the starting tree for each chain being the root model. The proposal distribution for the cross-chain swaps was uniform over the chain indexes $(2, ..., 10)$ and also uniform over the two implemented swap moves. On the top row, Figure 7 shows the unnormalised log posterior tree probability for the models visited by the mother chain, plotted respectively versus the iteration index and versus their number of leaves. Posterior sampling moved quickly towards areas of high marginal posterior probability models, which cluster the data over a range of 4 to 6 groups. The bottom plot in Figure 8 shows the estimated marginal inclusion probabilities for the nine covariates. The number of liver metastases, lobar involvement and a synchronous detection of the liver metastases along with the primary tumor appear to have prominent prognostic significance with respect to the remaining covariates, suggesting that the main determinant of survival for this sample are the extent of disease at the time of surgery and the accuracy of the diagnosis.

The estimated MAP tree clusters the 622 subjects into $\hat{\ell}_{MAP} = 4$ groups, respectively defined by the subsets

$$\begin{aligned}
\hat{\zeta}_1 &= \{NLM > 1\} \\
\hat{\zeta}_2 &= \{NLM \leq 1, DLM > 7\} \\
\hat{\zeta}_3 &= \{NLM \leq 1, DLM \leq 7, TD = 0\} \\
\hat{\zeta}_4 &= \{NLM \leq 1, DLM \leq 7, TD = 1\}
\end{aligned}$$

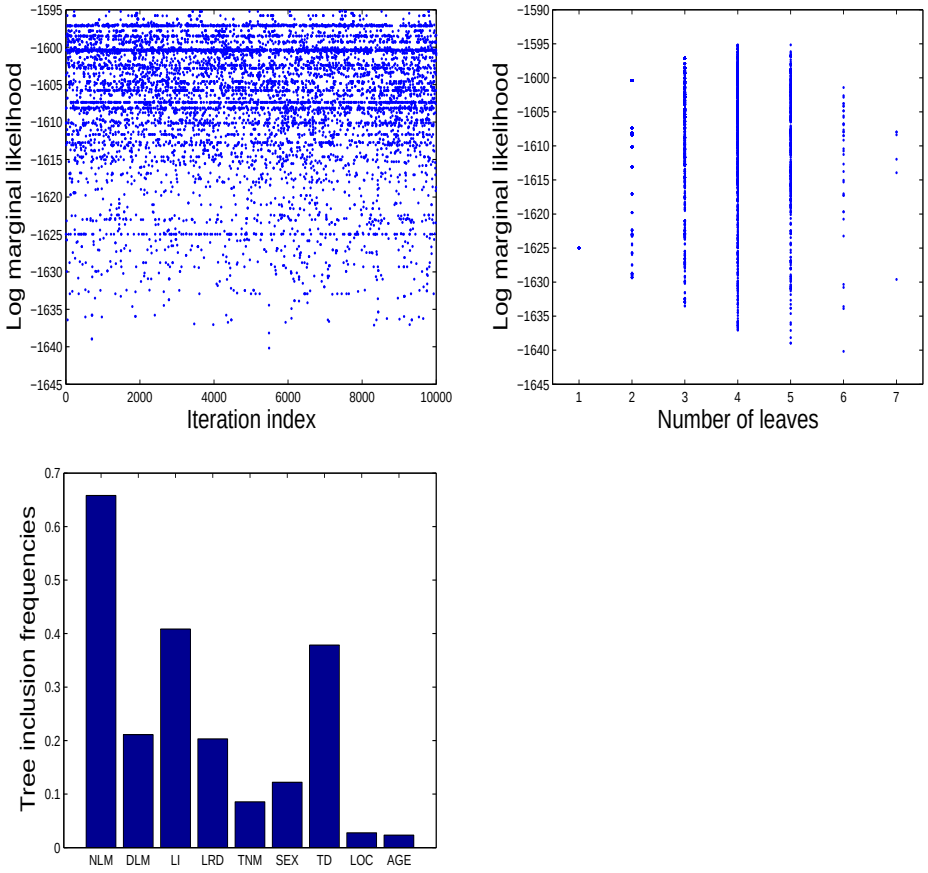

Figure 7: unnormalised log posterior tree probability for the models visited by the PHS mother chain plotted versus the sampler iteration index (top left) and the number of tree leaves (top right). PHS sampling using 9 auxilliary chains moves quickly towards areas of high marginal posterior probability models, which cluster the data over a range of 4 to 6 groups. Bottom plot: estimated marginal inclusion probabilities for the nine covariates. The number of liver metastases, lobar involvement and a synchronous detection of the liver metastases along with the primary tumor appear to have prominent prognostic significance with respect to the remaining covariates, suggesting that the main determinant of survival for this sample are the extent of disease at the time of surgery and the accuracy of the diagnosis.

Figure 9 shows that the estimated MAP tree separates the short-term survivors in leaf 1, who are characterised by a larger number of liver metastases of large size, from the long-term survivors in leaf 4, who present a few local metastases of small size without further symptoms. Leaves 2 and 3 represent intermediate survival scenarios characterised by either one metastasis of large diameter or by a late diagnosis of an originally limited metastatic process.

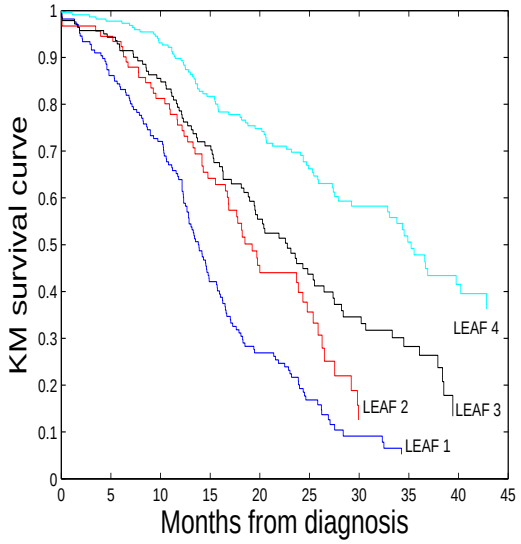

Figure 8: conditional Kaplan-Meier survival curves of the 622 colorectal cancer patients using the estimated MAP tree. The short-term survivors in leaf 1, who are characterised by a larger number of liver metastases of large size, are separated from the long-term survivors in leaf 4, who present a few local metastases of small size without further symptoms. Leaves 2 and 3 represent intermediate survival scenarios characterised by either one metastasis of large diameter or by a late diagnosis of an originally limited metastatic process.

5 Discussion

As noted by Geyer [1991], the attractive feature of multiple-chains MCMC samplers is that their target distribution factors into the product of the marginal distributions for each chain despite the fact that these chains are made dependent by the swap transitions. Under the standard conditions outlined in Section 2 we prove in Theorem 1 that samples generated by the PHS algorithm converge weakly to such product distribution. Furthermore, every time an auxilliary chain swaps with the mother chain its perfomance is improved in the sense of Peskun [1973]. This is reflected in the lower empirical autocorrelation of the auxilliary chain and lower integrated autocorrelation times reported in Figure 2.

In Section 2 we also noted that the joint transition kernels of PT, PHS and sPHS are mixtures where $(M-1)$ out of the M parallel chains are auxilliary to the update of the first chain. The complexity of these transition kernels, has so far hindered a direct analytical comparison of their convergence properties. Establishing orderings between the two kernels using the criteria illustrated in Peskun [1973], Meyn and Tweedie [1994] and Mira [2001] is thus object of ongoing research. Also, in light of the empirical measures of accuracy reported in Table 2 an other very promising topic in this area is the formulation of adaptive multiple-chains sampling strategies (Craiu and Meng [2005], Craiu et al. [2009]).

In the reminder of this section we briefly discuss the main analogies and differences between PHS and its most closely related algorithms in the current literature. In section 2 we noted that by construction PHS produces a mother chain which always moves but which exhibits low serial dependence. This property marks the most evident difference between the sample

paths of the first chain of PHS, those of the MH algorithm and those of the cold chain in PT. With respect to the latter, PHS focuses on using simultaneously many different proposal distributions as opposed to adopting tempered target distributions. This is a key feature for the applicability of parallel samplers for Bayesian inference as one does not need worring about which temperature values correspond to proper tempered target posterior distributions.

The PHS algorithm is also strictly related to at least two other categories of MCMC samplers: those using latent variables and multiple-try proposals. Parameter augmentation was introduced in the context of single-chain MCMC samplers by Tanner and Wong [1987] and later by Neal [2000] to improve mixing of a chain by expanding its state-space using additional dimensions. Conditionally on these auxilliary variables the posterior distribution of the parameters of interest can typically be sampled exactly via a Gibbs step. The PHS algorithm can be seen as a variable augmentation scheme where the auxilliary coefficients are $M-1$ replicates of the parameter of interest itself. The main analogy between the algorithm of Liu et al. [2000] and PHS is that many candidates are available at each iteration to update one chain of interest. In Liu et al. [2000], only one of such updates is retained and the Metropolis ratio is modified accordingly. In PHS the proposal mechanism generating all potential updates is not constrained to be the same for all chains and all values not used for swapping with the mother chain are used for updating the other $M-1$ parallel chains.

References

A. Antoniadis, G. Grégoire, and G. Nason. Density and hazard rate estimation for right-censored data by using wavelet methods. *Journal of the Royal Statistical Society B*, 61:63–84, 1999.

Y.F. Atchadé and J.S. Rosenthal. On adaptive Markov chain Monte Carlo algorithms. *Bernoulli*, 11:815–828, 2005.

K.B. Athreya, H. Doss, and J. Sethuraman. On the convergence of the Markov Chain simulation method. *The Annals of Statistics*, 24:69–100, 1996.

Y. Bai. Simultaneous drift conditions for adaptive Markov chain Monte Carlo algorithms. *University of Toronto Preprint*, 2009.

Y. Bai, G.O. Roberts, and J.S. Rosenthal. On the containment condition for adaptive Markov chain Monte Carlo algorithms. *University of Toronto Preprint*, 2009.

M.M. Barbieri and J.O. Berger. Optimal predictive model selection. *The Annals of Statistics*, 3:870–897, 2004.

M. Bédard and J.S. Rosenthal. Optimal scaling of Metropolis algorithms: Heading towards general target distributions. *The Canadian Journal of Statistics*, 36: 483–502, 2008.

F. Berrino, A. Verdecchia, J.M. Lutz, C. Lombardo, A. Micheli, and R. Capocaccia. Comparative cancer survival information in Europe. *European Journal of Cancer*, 45:901–908, 2009.

L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

L. Breiman. Population theory for boosting ensembles. *The Annals of Statistics*, 32:1–11, 2004.

L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Chapman and Hall, New York, 1984.

A.E. Brockwell and J.B. Kadane. Identification of regeneration times in MCMC simulation, with applications to adaptive schemes. *Journal of Computational and Graphical Statistics*, 14:436–458, 2005.

S. P. Brooks and G.O. Roberts. Assessing convergence of Markov chain Monte Carlo algorithms. *Statistics and Computing*, 8:319–335, 1998.

S.P. Brooks. MCMC convergence diagnosis via ultivariate bounds on log-concave densities. *The Annals of Statistics*, 26:398–433, 1998.

S.P. Brooks, P. Giudici, and G.O. Roberts. Efficient construction of Reversible Jump MCMC proposal distributions. *Journal of the Royal Statistical Society B*, 65:3–55, 2003.

O. Cappé and C.P. Robert. Markov chain Monte Carlo: 10 years and still running! *Journal of the American Statistical Association*, 95:1282–1286, 2000.

O. Cappé, G.O. Roberts, and T. Ryden. Reversible jump, birth-and-death and more general continuous time Markov chain Monte Carlo samplers. *Journal of the Royal Statistical Society B*, 65:679–700, 2003.

B.P. Carlin and S. Chib. Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society B*, 57, 1995.

G. Celeux, M. Hurn, and C.P. Robert. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95:957–970, 2000.

D. Chauveau and P. Vandekerkhove. Improving convergence of the Hastings-Metropolis algorithm with a learning proposal. *Scandinavian Journal of Statistics*, 29:13–29, 2002.

Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. Bayesian CART model search. *Journal of the American Statistical Society*, 93:935–947, 1998.

M. Clyde and E.I. George. Model Uncertainty. *Statistical Science*, 19:81–94, 2004.

J. Corander, M. Gyllenberg, and T. Koski. Bayesian model learning based on a parallel MCMC strategy. *Statistics and Computing*, 16:355–362, 2006.

M.K. Cowles and B.P. Carlin. Markov chain Monte Carlo algorithms convergence diagnostics: a comparative reiview. *Journal of the American Statistical Association*, 91:883–904, 1996.

R.V. Craiu and X.L. Meng. Multi-process parallel antithetic coupling for forward and backward Markov chain Monte Carlo. *The Annals of Statistics*, 33:661–697, 2005.

R.V. Craiu, J.S. Rosenthal, and C. Yang. Learn from thy neighbor: Parallel-chain adaptive MCMC. *Journal of the American Statistical Association, To Appear*, 2009.

R.B. Davis and J.R. Anderson. Exponential survival trees. *Statistics in Medicine*, 8, 1989.

P. Dellaportas, J.J. Forster, and I. Ntzoufras. On Bayesian model and variable selection using MCMC. *Statistics and Computing*, 12:27–36, 2002.

D.G.T. Denison, B.K. Mallick, and A.F.M. Smith. A Bayesian CART algorithm. *Biometrika*, 85, 1998.

S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195:216–222, 1987.

J.M Flegal, M. Haran, and J.L. Jones. Markov chain Monte Carlo: can we trust the third significant figure? *Statistical Science*, 23:250–260, 2008.

D. Gamerman. *Markov chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman and Hall, 1997.

J. Gasemyr. On an adaptive version of the Metropolis-Hastings algorithm with independent proposal distribution. *Scandinavian Journal of Statistics*, 30:159–173, 2003.

A.E. Gelfand and S.K. Sahu. On Markov chain Monte Carlo acceleration. *Journal of Computational and Graphical Statistics*, 3:261–276, 1994.

A.E. Gelfand and A. F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Society*, 85:398–409, 1990.

E.I. George and R.E. McCulloch. Approaches for Bayesian variable selection. *Statistica Sinica*, 7:339–373, 1997.

E.I. George and R.E. McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88:882–889, 1993.

J. Geweke. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. *In: Bayesian Statistics 4, J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith editors; Oxford Press*, pages 169–194, 1992.

C. J. Geyer and E.A. Thompson. Annealing Markov Chain Monte Carlo with applications to ancestral inference. *Journal of the American Statistical Association*, 90:909–920, 1995.

C.J. Geyer. Markov chain Monte Carlo maximum likelihood. *Computing Science and Statistics: Proceedings on the 23rd Symposium on the Interface, New York*, 1991.

W.R. Gilks and P. Wild. Adaptive rejection sampling for Gibbs sampling. *Journal of the Royal Statistical Society C*, 41:337–348, 1992.

W.R. Gilks, G.O. Roberts, and E.I. George. Adaptive direction sampling. *The Statistician*, 43:179–189, 1994.

W.R. Gilks, N.G. Best, and K.K.C. Tan. Adaptive rejection Metropolis sampling within Gibbs sampling. *Applied Statistics*, 44:455–472, 1995a.

W.R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov chain Monte Carlo in practice*. Chapman and Hall, 1995b.

W.R. Gilks, N.G. Best, and K.K.C. Tan. Adaptive Markov chain Monte Carlo through regeneration. *Journal of the American Statistical Association*, 93:1045–1054, 1998.

J. Gill and G. Casella. Dynamic tempered transitions for exploring multimodal posterior distributions. *Political Analysis*, 12:425–443, 2004.

L. Gordon and R.A. Olshen. Tree-structured survival analysis. *Cancer Treatment Reports*, 69, 1995.

P.H. Green. Reversible Jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.

P.J. Green and X.L Han. Metropolis methods, Gaussian proposals and antithetic variables. *Lecture notes in Statistics*, 74:142–164, 1991.

P.J. Green and A. Mira. Delayed rejection in reversible jump Metropolis Hastings. *Biometrika*, 88:1035–1053, 2001.

M.P.W. Grocott, D.S. Martin, D.Z.H Levett, H. Montgomery, and M.G. Mythen. Caudwell Xtreme Everest. *Anaesthesia and Analegesia*, 6:81–84, 2008.

H. Haario, E. Saksman, and J. Tamminen. Adaptive proposal distribution for random walk metropolis algorithm, 1999.

H. Haario, E. Saksman, and J. Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7:223–242, 2001.

H. Haario, E. Saksman, and J. Tamminen. Component-wise adaptatation for high-dimensional MCMC. *Computational Statistics*, 20:265–273, 2005.

H. Haario, M Laine, A. Mira, and E. Saksman. DRAM: efficient adaptive MCMC. *Statistics and Computing*, 16:339–354, 2006.

U.H.E. Hansmann. Parallel tempering algorithm for conformational studies of biological molecules. *Chem. Phys. Lett.*, 281:140–152, 1997.

G. Haupt and U. Mansmann. Survival trees in Splus. *Advances in Statistical Software 5*, pages 615–622, 1995.

P. Hermanek and F.P. Gall. Uicc Studie zur Klassifikation von Lebermetastasen. *Chirurgie der Lebermetastasen und primaren malignen Tumoren*, 1990.

T.K. Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:832 – 844, 1998.

B. Hu and K-W. Tsui. Distributed evolutionary Monte Carlo for Bayesian computing. *Computational Statistics and Data Analysis*, doi:10.1016/j.csda.2008.10.025, 2008.

J.P. Huelsembeck, B. Larget, R.E. Miller, and F. Ronquist. Potential applications and pitfalls of Bayesian inference of phylogeny. *Syst. Biol.*, 51:673–688, 2002.

J.P. Huelsembeck, B. Larget, and M.E. Alfaro. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. *Mol. Biol. Evol.*, 21: 1123–1133, 2004.

J.P. Huelsenbeck and F. Ronquist. Bayesian analysis of molecular evolution using MrBayes. *Statistical methods in molecular evolution*, 2:183–226, 2005.

K. Hukushima and K. Nemoto. Exchange Monte Carlo method and application to Spin Glass simulations. *Journal of the Physical Society of Japan*, 65:1604–1620, 1996.

D. Husmeier and G. McGuire. Detecting recombination in 4-taxa DNA sequence alignments with Bayesian hidden Markov models and Markov chain Monte Carlo. *Mol. Biol. Evol.*, 20:315–337, 2003.

Y. Iba. Extended ensemble Monte Carlo. *International Journal of Modern Physics*, 12:623–656, 2001.

S.F. Jarner and G.O. Roberts. Polynomial convergence rates of Markov chains. *The Annals of Applied Probability*, 12:224–247, 2002.

A. Jasra, D.A. Stephens, and C.C. Holmes. Population-based reversible jump Markov chain Monte Carlo. *Biometrika*, 94:787–807, 2007.

A. Kottas. Nonparametric Bayesian survival analysis using mixtures of Weibull distributions. *Journal of Statistical Planning and Inference*, 136:578–596, 2006.

L. Kuo and B. Mallick. Variable selection for regression models. *Sankhya B*, 60: 65–81, 1998.

C. Lakner, P. Van der Mark, J.P. Huelsenbeck, B. Larget, and F. Ronquist. Efficiency of Markov chain Monte Carlo tree proposals in Bayesian phylogenetics. *Syst. Biol.*, 57:86–103, 2008.

M. Li. *Bayesian discovery of regulatory motifs using reversible jump Markov chain Monte Carlo*. PhD thesis, University of Washington, 2006.

S. Li, D.K. Pearl, and H. Doss. Phylogenetic tree construction using Markov chain Monte Carlo. *Journal of the American Statistical Association*, 95:493–508, 2000.

F. Liang and W.H. Wong. Real-parameter evolutionary Monte Carlo with applications to Bayesian mixture models. *Journal of the American Statistical Association*, 96:653–666, 2001.

C.Y. Lin, C.H. Hu, and U.H.E. Hansmann. Parallel tempering simulations of HP-36. *Proteins: Structure, Functions and Genetics*, 53:436–445, 2003.

J.S. Liu. *Monte Carlo Strategies in Scientific computing*. Springer, 2001.

J.S. Liu and C. Sabatti. Simulated sintering: Markov chain Monte Carlo with spaces of varying dimensions. *Bayesian Statistics*, 6:389–403, 1998.

J.S. Liu, F. Liang, and W.H. Wong. Dynamic weighting in Monte Carlo and optimization. *Journal of the American Statistical Association*, 94:14220–14224, 1997.

J.S. Liu, F. Liang, and W.H. Wong. The multiple-try method and local optimization in Metropolis sampling. *Journal of the American Statistical Association*, 95:121–134, 2000.

J.S. Liu, F. Liang, and W.H. Wong. A theory for dynamic weighting in Monte Carlo computation. *Proceedings of the National Academy of Sciences*, 96:561–573, 2001.

D.J. Lunn, A. Thomas, N. Best, and D. Spiegelhalter. WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10:325–337, 2000.

G. Lunter, I. Miklos, A. Drummond, J.L. Jensen, and J. Hein. Bayesian coestimation of phylogeny and sequence alignment. *BMC Bioinformatics*, 6:83–93, 2005.

B. Mau, M.A. Newton, and B. Larget. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics*, 55:1–12, 1999.

I.W. McKeague and W. Wefelmeyer. Markov chain Monte Carlo and Rao-Blackwellization. *Journal of Statistical Planning and Inference*, 85:171–182, 2000.

K.L. Mengersen and R.L. Tweedie. Rates of convergence of the Hastings and Metropolis algorithms. *The Annals of Statistics*, 24:101–121, 1996.

N. Metropolis and S. Ulam. The Monte Carlo method. *Journal of the American Statistical Association*, 44:335–341, 1949.

N. Metropolis, A. Rosenbluth, M. Rosenbluth, M. Teller, and E. Teller. Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–2092, 1953.

S. Meyn and R.L. Tweedie. State-dependent criteria for convergence of Markov chains. *The Annals of Applied Probability*, 4:149–168, 1994.

A. Mira. Ordering and improving the performance of Monte Carlo Markov chains. *Statistical Science*, 16:340–350, 2001.

A. Mira and C. Geyer. Ordering Monte Carlo Markov chains. *Technical Report 632, School of Statistics, University of Minnesota*, 1999.

A. Mira and D.J. Sargent. A new strategy for speeding Markov chain Monte Carlo algorithms. *Statistical Methods and Applications*, 12:49–60, 2003.

T.J. Mitchell and J.J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83:1023–1032, 1988.

M.Leblanch and J.Crowley. Relative risk trees for censored survival data. *Biometrics*, 48, 1992a.

M.Leblanch and J.Crowley. Survival trees by goodness of split. *Journal of the American Statistical Association*, 88, 1992b.

E. Mossel and E. Vigoda. Phylogenetic MCMC algorithms are misleading on mixtures of trees. *Sience*, 309:2207–2209, 2005.

E. Mossel and E. Vigoda. Limitations of Markov chain Monte Carlo algorithms for inference of phylogeny. *The Annals of Applied Probability*, 16:2215–2234, 2006.

J.W. Myers and K.B. Laskey. Population Markov chain Monte Carlo. *Machine Learning*, 50:175–196, 2001.

P. Neal and G.O. Roberts. Optimal scaling for partially updating MCMC algorithms. *The Annals of Applied Probability*, 16:475–515, 2006.

P. Neal, G.O. Roberts, and J. Yuen. Optimal scaling of of random walk Metropolis algorithms with discontinous target densities. *University of Manchester Technical Report*, 1, 2007.

R. Neal. Slice sampling. *Technical report No. 2005 - Dept. of Statistics, University of Toronto*, 2000.

R.M. Neal. Sampling from multimodal distributions using tempered transitions. *Statistics and Computing*, 6:353–366, 1996.

R.M. Neal. Probabilistic inference using Markov chain Monte Carlo methods. *Technical report, Department of Computer Science, University of Toronto*, 1993.

D.J. Nott and P.J. Green. Bayesian veriable selection and the Swendsen-Wang algorithm. *Journal of Computational and Graphical Statistics*, 13:141–157, 2004.

E. Nummelin. *General Irreducible Markov Chains on Non-Negative Operators*. Cambridge University Press, 1984.

L.M. Pasetto, E. Rossi, and S. Monfardini. Liver metastases of colorectal cancer: medical treatment. *Anticancer*, 23:4245–56, 2003.

P.H. Peskun. Optimum Monte Carlo sampling using Markov chain. *Biometrika*, 60:607–612, 1973.

G. Petris and L. Tardella. A geometric approach to transdimensional Markov chain Monte Carlo. *The Canadian Journal of Statistics*, 31:469–482, 2003.

J. Pittman, E. Huang, H. Dressman, C.F. Horng, S.H. Cheng, M.H. Tsou, C.M. Chen, A. Bild, E.S. Iversen, A.T. Huang, J.R. Nevins, and M. West. Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *Proceedings of the National Academy of Sciences*, 101:8431–8436, 2004.

A. E. Raftery, D. Madigan, and J. A. Hoeting. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92:179–191, 1997.

R. Ren and G. Orkoulas. Parallel Markov chain Monte Carlo simulations. *Journal of Chemical Physics*, doi:10.1063/1.2743003, 2007.

C.P. Robert. Convergence control methods for Markov chain Monte Carlo algorithms. *Statistical Science*, 10:231–253, 1995.

C.P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 1999.

G.O. Roberts and J.S. Rosenthal. Optimal scaling for various Metropolis Hastings algorithms. *Statistical Science*, 16:351–367, 2001.

G.O. Roberts and J.S. Rosenthal. Coupling and ergodicity of adaptive MCMC. *Journal of Applied Probability*, 44:458–475, 2007.

G.O. Roberts and J.S. Rosenthal. Examples of adaptive MCMC. *University of Toronto Preprint*, 2008.

G.O. Roberts and J.S. Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society B*, 60:255–268, 1998a.

G.O. Roberts and J.S. Rosenthal. Markov chain Monte Carlo: some practical implementations of theoretical results. *The Canadian Journal of Statistics*, 26: 5–20, 1998b.

G.O. Roberts and O. Stramer. Tempered Langevin diffusions and algorithms. *University of Iowa, Department of Statistics and Actuarial Sciences Technical Report*, 314, 2002.

G.O. Roberts and R.L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2:341–363, 1996a.

G.O. Roberts and R.L. Tweedie. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83:95–110, 1996b.

G.O. Roberts, A. Gelman, and W.R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7: 110–120, 1997.

J.S. Rosenthal. Parallel computing and Monte Carlo algorithms. *Far East Journal of Theoretical Statistics*, 4:207–236, 2000.

J.S. Rosenthal. Convergence rates for Markov chains. *SIAM review*, 37:387–405, 1995.

M. Sant, C. Allemani, M. Santaquilani, A. Knijn, F. Marchesi, and Capocaccia R. EUROCARE-4. Survival of cancer patients diagnosed in 1995-1999. Results and commentary. *European Journal of Cancer*, 45:931–991, 2009.

S.A. Sission. Transdimensional Markov chains. *Journal of the American Statistical Association*, 100:1077–1089, 2005.

A. F.M. Smith and G.O. Roberts. Bayesian computations via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society B*, 55:3–23, 1993.

M. Smith and R. Kohn. Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, 75:317–343, 1996.

A.D. Sokal. Monte Carlo methods in Statistical Mechanics: Foundations and new algorithms. *Lecture Notes at the Cargese summer school on "Functional Integration: basis and applications"*, 1996.

M. Stephens. Bayesian analysis of mixture models with an unknown number of components - an alternative to reversible jump methods. *Annals of Statistics*, 28:40–74, 2000.

R.H. Swedensen and J.S. Wang. Nonuniversal critical dynamics in Monte Carlo simulations. *Physical Review Letters*, 58:86–88, 1987.

M.A. Tanner and W.H. Wong. The calculation of posterior distributions via data augmentation. *Journal of the American Statistical Association*, 82:528–541, 1987.

L. Tierney. Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22:1701–1728, 1994.

L. Tierney and A. Mira. Some adaptive Monte Carlo methods for Bayesian inference. *Statistics in Medicine*, 18:2507 – 2515, 1991.

R. Waagepetersen, N. Ibanez-Escriche, and D. Sorensen. A comparison of strategies for Markov chain Monte Carlo in quantitative genetics. *Genet. Set. Evol.*, 40: 161–176, 2008.

M. West. Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian Statistics 3*, 2003.

U. Wolff. Monte Carlo errors with less errors. *Computer Physics Communications*, 156:143–153, 2004.

D.B. Woodard, S.C. Schmidler, and Huber M.L. Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions. *Annals of Applied Probability*, 19:617–640, 2009.

Z. Yang and B. Rannala. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. *Mol. Biol. Evol.*, 14:717–724, 1997.

T. Yoshida, M. Chida, M. Ichioka, and Y. Suda. Blood lactate parameters related to aerobic capacity and endurance performance. *European Journal of Applied Physiology and Occupational Physiology*, 56:7–11, 1987.

Z. Zheng. On swapping and simulated tempering algorithms. *Stochastic processes and their applications*, 104:131–154, 2001.

Acknowledgements

The first author acknowledges the partial support received through a Visiting Fellow grant of the University of Insubria during the development of this work. A. Mira acknowledges the partial support of the PRIN (Italian National Research Program) $2007XECZ7L\_003$. The software implementing the models and the MCMC algorithms employed in Sections 3, 4 and 5 can be obtained in the form of `MATLAB` modules upon request to the first author.